Journal of Physics: Conference Series **119** (2008) 042015 doi:10.1088/1742-6596/119/4/042015

Experience and Lessons learnt from running High Availability Databases on Network Attached Storage

Manuel Guijarro, Ruben Gaspar et al CERN IT/DES

CERN IT-DES, CH-1211 Geneva 23, Switzerland

Manuel.Guijarro@cern.ch, Ruben.Gaspar.Aparicio@cern.ch

Abstract. The Database and Engineering Services Group of CERN's Information Technology Department supplies the Oracle Central Database services used in many activities at CERN. In order to provide High Availability and ease management for those services, a NAS (Network Attached Storage) based infrastructure has been setup. It runs several instances of the Oracle RAC (Real Application Cluster) using NFS (Network File System) as shared disk space for RAC purposes and Data hosting. It is composed of two private LANs (Local Area Network), one to provide access to the NAS filers and a second to implement the Oracle RAC private interconnect, both using Network Bonding. NAS filers are configured in partnership to prevent having single points of failure and to provide automatic NAS filer fail-over.

1. Introduction

This paper describes a NAS based infrastructure and its implementation using a Fabric Management framework such as the Quattor administration toolkit. It also covers aspects related to NAS performance and monitoring as well as Data Backup and Archive of such facility using already existing infrastructure at CERN.

A NAS is the name given to dedicated Data Storage technology that can be connected directly to a Computer Network to provide centralized data access and storage to heterogeneous network clients. Operating System and other software on the NAS unit provide only the functionality of data storage, data access and the management of these functionalities. Several file transfer protocols are supported (NFS, SMB, etc).

By contrast to a SAN (Storage Area Network), NAS uses file-based protocols where it is clear that the storage is remote, and Server nodes request a portion of an abstract file rather than a disk block. SAN is an architecture to attach remote computer storage devices such as disk arrays, tape libraries, etc to servers in such a way that, to the Operating System, the devices appear as locally attached devices. SANs tend to be expensive and complex which makes them uncommon outside larger enterprises. Moreover, when using SAN for Oracle Databases, it's recommended to use Oracle ASM (Automatic Storage Management). ASM adds an additional layer of complexity.

The main motivation to setup NAS based infrastructure for Oracle Database is:

© 2008 IOP Publishing Ltd

- To provide the file sharing needed for Oracle RAC. NAS storage presents some advantages like the possibility of having remote storage or the fact that communication can be based on standard protocols such as NFS.
- To ease relocation of services within Server nodes
- To use NAS specific features: Snapshots, RAID, Failover based on NAS partnership, Dynamic re-Sizing of file systems, Remote Sync to offsite NAS, etc
- To use Ethernet rather than Fiber Channel (which is more difficult to manage as the protocol is more complex per-se and extra parameterization is required like defining zones, etc.)
- To ease File Server Management: automatic failure reporting to vendor, etc
- To simplify administration of Database storage (scalability, easiness to add/create/resize NAS volumes, etc.)

2. Topology

The topology is composed of several (HP DL 380 G5) Server nodes running Red Hat Enterprise Linux 4. Each of them is equipped with two dual core 2.33 GHz Intel processors, 8 GB RAM and 6 NICs (Network Interface Cards). Oracle RAC requires a private network interface between RAC members. This is implemented by connecting all server nodes to 2 network switches. The 2 NICs connected to those switches are aggregated using Linux Bonding and are seen as a single network interface (with a single IP address).

In a similar way, each server node is connected to 2 network switches that reach the NAS filers that are connected to two Disk Shelves. A fifth NIC of the server is connected to the GPN (General Purpose Network). In this topology there is no single point of failure other than the switch which connects all servers to the GPN. A failure of a server node is overcome by the Oracle RAC software that redirects requests to another Server node within the same RAC.

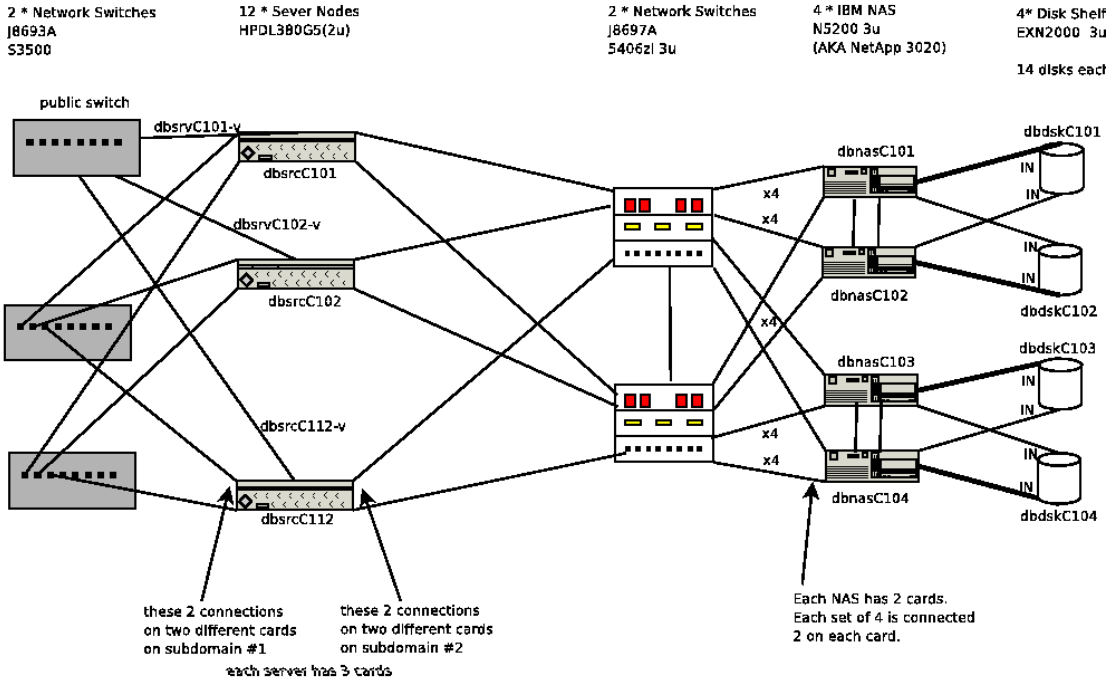

Figure 1. NAS based infrastructure topology.

2.1. Bonding

Network bonding (also known as port trunking) consists of aggregating multiple network interfaces into a single logical bonded interface that correspond to a single IP address. This technique allows implementation of load balancing (i.e. using multiple network ports in parallel to increase the link speed beyond the limits of a single port) and/or automatic fail-over (in the even of a network interface failure, data is transferred through other network interfaces in the same Bonding aggregate). Two types of bonding modes are used in this topology:

- Active-backup mode: Only one NIC in the logical bonded interface is active. A different NIC becomes active if, and only if, the active NIC fails. The bond's MAC address is externally visible on only one port (network adapter) to avoid confusing the switch. The "primary" bonding option can be used to specify which NIC is the primary device (for example: *"primary" = eth3*). The specified device will always be the active NIC while it is available. Only when the primary device is off-line will alternate devices be used. This is useful when one NIC is preferred over another, e.g., when one NIC has higher throughput than another. The link state of each NIC is monitored every 100 milliseconds.

- IEEE 802.3ad Dynamic link aggregation mode: For trunking between NAS filers and switches as well as for the connection between the switches. Each NAS filer is equipped with 8 NICs that are aggregated on 2 VIFs (Virtual Interfaces) of 4 NICs each. Each filer does load balancing on the network traffic transmitted over each VIF. This is done using the IP based method, i.e.: the outgoing interface is selected on the basis of the NAS filer and client's IP address. It is also possible to use MAC based or Round-Robin methods instead.

2.2. NetApp Cluster Configuration

Each NAS filer is configured in NetApp Cluster enabled mode, having as a partner another NAS filer which accesses the same set of Disk Shelves and the same subnet. NAS filers use a cluster interconnects to monitor each other. When a NAS filer fails, a take over occurs, and the partner filer continues to serve the failed filer's data. Take over operations can be manually initiated. This allows performing non-disruptive NAS filer software upgrades as well as disk storage maintenance.

The 2 VIFs in each NAS filer are connected to different network switches. They are aggregated in a 2nd level VIF which acts in active-backup mode, i.e. only one of the 1st level VIFs is active.

3. NAS filer and Disk Shelves setup

Each pair of NAS filers is connected to a shared set of Disk Shelves via FC (Fiber Channel). Each shelf is equipped with 14* 146 GB disks (or 14* 300 GB disks). Data ONTAP 7.2.2 is the operating system used in all NAS filers.

3.1. Disk Aggregates and Volume

A single aggregate is created in each shell containing all disks in it (13 disks + 1 spare disk). It is composed of a single RAID-DP (RAID Double-Parity). RAID-DP is NetApps' implementation of RAID 6 (provides fault tolerance from two drive failures) that uses double parity for data protection.

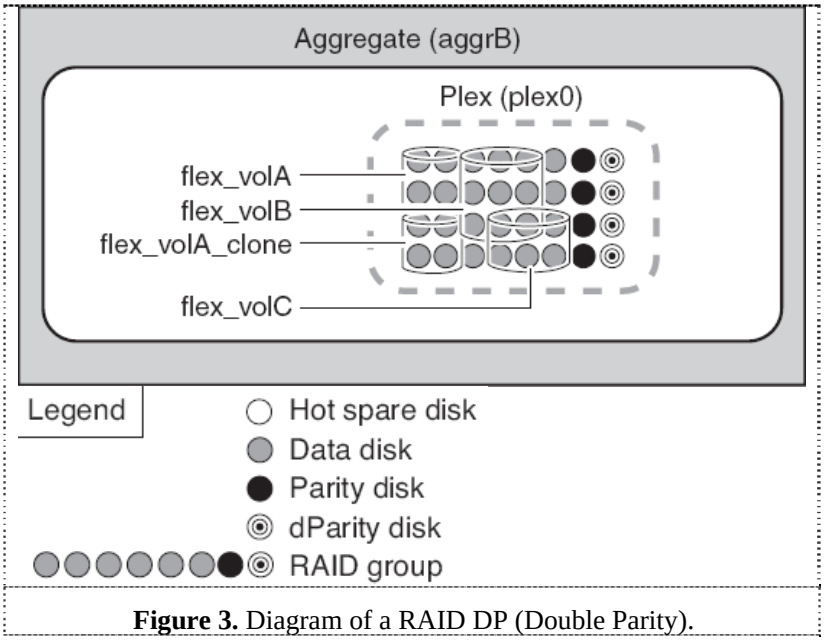

Figure 3. Diagram of a RAID DP (Double Parity).

RAID-DP improves standard RAID 6 performance due to the behavior of the storage controller software. All file system requests are first written to the battery backed NVRAM to ensure there is no data loss should the system lose power. Blocks are never updated in place, so when incoming write

operations are performed, writes are aggregated and the storage controller tries to write only complete stripes including both parity blocks. RAID-DP provides better protection than RAID1/0, and even enables disk firmware updates to occur in real-time without any outage.

Each aggregate contains several FlexVol volumes (sometimes called Flexible Volume). These are loosely coupled to its containing aggregate, as opposed to traditional volumes where each aggregate contained a single volume. Since the FlexVol volume is managed separately from the aggregate, it is possible to create small FlexVol volumes (20 MB or larger), and increase or decrease their size in increments as small as 4 KB.

A FlexVol volume can share its containing aggregate with other FlexVol volumes. Thus, a single aggregate can be the shared source of all the storage used by all the FlexVol volumes contained by that aggregate.

Each FlexVol volume is created with a small size and the autosize option, which allows automatic increase of volume size as this one grows.

3.2. NFS access options

To NFS mount options used to mount the NAS filer volumes where Server nodes write Oracle Data are those suggested by the vendor in this kind of RAC configurations. These are:

mount –o rw,bg,hard,nointr,tcp,vers=3,actimeo=0,timeo=600,rsize=32768,wsize=32768

Using the "*bg"* option means that a Server node will be able to finish booting without waiting for any NAS Filer. The "*hard"* option minimizes the likelihood of data loss during network and server instability, while "*nointr"* doesn't allow file operations to be interrupt. The "*tcp"* option forces NFS to use TCP protocol and works well on many typical LANs with 32KB read and write size. The "*timeo"* option is used to RPC retransmission timeouts. Retransmission is the mechanism by which clients ensure a server receives and processes an RPC request. If the client does not receive a reply for an RPC within a certain interval for any reason, it retransmits the request until it receives a reply from the server. After each retransmission, the client doubles the retransmit timeout up to 60 seconds to keep network load to a minimum. Using *"timeo=600"* is a good default for TCP mounts.

3.3. Backup

All Oracle Databases hosted in NAS filers as well as internal disks of server nodes are backed up in CERN's central network backup system, TSM (IBM Tivoli Storage Manager). RMAN (Oracle Recovery Manager) is used to backup up Oracle Databases using TDPO (TSM Data Protection for Oracle) client. TSM and TDPO software are installed and configured using Quattor components.

Data ONTAP provides Snapshots of volumes, which are a read-only copy of an entire volume that protects against accidental deletions or modifications of files without duplicating file contents. This feature is used for backup purposes before Database maintenance operations and it is typically preceded by a Database shutdown.

4. Performance

NetApps provides SAN as well as NAS based solutions. In their Performance Report (http://www.netapp.com/library/tr/3423.pdf), it is stated that the throughput with NFS NAS based solutions is slightly lower than that of iSCSI (Internet SCSI protocol) or FCP (Fiber Channel Protocol) based solutions.

Oracle Database performs very well with the NAS devices that had been tested and put into production. Many tests were run from the Database, in single instance mode and in cluster (RAC) mode.

4.1. Direct I/O

Oracle makes use of direct I/O with NFS to access Database files stored on Network Attached Storage devices. Direct I/O avoids external caching in the OS page cache. Moreover, it is much more performing (for typical Oracle I/O Database workloads) than buffered I/O. It was tested that performance doubled when going from buffered I/O to direct I/O. The only necessary required step to enable direct I/O at the Database level is to have the FILESYSTEMIO_OPTIONS parameter set to direct IO:

alter system set FILESYSTEMIO_OPTIONS = directIO scope=spfile sid='';*

4.2. Measurements

A stress test program has been written that simulates the random "small I/O" (typically 8KB read operations) performed by Oracle Database. All tests show that each disk in a NAS (test performed with a set of aggregates summing up a total of 12 disks) provides a maximum of about 150 I/O operations per second and per device (with 32 threads, see Figure bellow). This equals to the maximum of I/O operations that can be performed in a single disk when doing random operations. The disks used were of 10000-RPM Fibre Channel type.

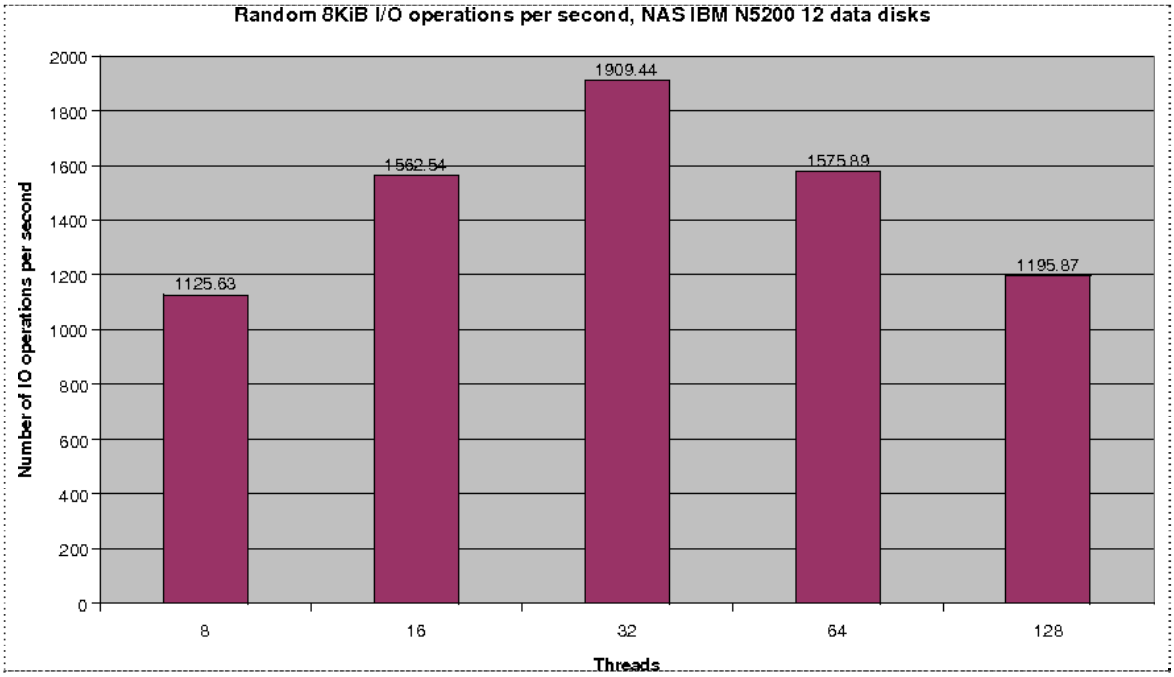

Figure 3. Operations per second/Threads

With more than 64 threads disks start saturating and the number of operations starts dropping.

5. Software Installation and Customization

All Server nodes are Quattor (http://cern.ch/quattor) managed. Quattor is a system administration toolkit providing a powerful, portable and modular toolsuite for the automated installation, configuration and management of clusters and farms running UNIX derivates like Linux and Solaris. All information regarding Server nodes' software setup is stored in a central CDB (Configuration Database). The set of already available Quattor NCM components is enough to fully automate configuration (including setup for Network Bonding, NFS, Firewall, Kernel Modules, etc) for everything regarding the Operating System layer. On the other hand, a lot of work has been done to automate installation and configuration of Oracle RDBMS and RAC software.

5.1. Standard Database setup

For each Oracle RAC Database, four volumes are created in the NAS filer. Critical files are spread across those volumes. They are mounted in each RAC member on four mount points with the following content:

- /ORA/dbs00: logging directory structure and a copy of the control file
- /ORA/dbs02: copy of the control file, copy of the voting disk and the archive redo files
- /ORA/dbs03: the spfile, a copy of the control file, a copy of the voting disk, datafiles and a copy of the registry
- /ORA/dbs04: a copy of the control file, a copy of the voting disk, a copy of the registry. This volume is located in a different aggregate, i.e. a different NAS filer and Disk Shelf.

5.2. Oracle Software Installation

The Oracle Clusterware and RDBMS software were packaged in an RPM format. Files included in the RPM are obtained using OUI (Oracle Universal Installer) for a first software installation and Oracle "cloning" script is used for distribution on the different RAC nodes. Due to Oracle installation procedures restrictions, these RPMs are only used for software installation and cannot serve for security checks or dependency verifications. Elapsed time for the fully automated RPM installation is a few minutes and can be compared to a good fraction of an hour using interactive Oracle tools. Package Removal (AKA "Uninstallation" in Oracle terminology) is also much easier.

The work required to build RPMs starting from OUI is not negligible and it is comparable to a single manual interactive installation including applying of all patch sets as well as isolated patches.

5.3. Standard RAC Configuration

The Quattor CDB contains all information about what should be installed and configured on every RAC member as well as the configuration information for of each RAC instance. A Quattor component reads the CDB to obtain all this configuration information and then it modifies RAC configuration files and start-up procedures accordingly. Finally, it starts the cluster ware and the RDBMS.

6. Monitoring

The Lemon Monitoring System (http://cern.ch/lemon) is used to monitor all aspects of the Server Nodes including resource usage, system errors, etc. No particular Lemon sensor has been developed for this infrastructure since the already existing wide set of sensors covers most of what is needed. Lemon is a server/client based monitoring system. On every monitored node, a monitoring agent launches and communicates using a push/pull protocol with sensors that are responsible for retrieving monitoring information. The extracted samples are stored on a local cache and forwarded to a central Measurement Repository using UDP or TCP transport protocol with or without authentication/encryption of data samples. Sensors can collect information on behalf of remote entities like switches or power supplies. The Measurement Repository can interface to a relational Database or a flat-file backend for storing the received samples. Web based interface is provided for visualizing the data

6.1. NAS Monitoring

NAS filers cannot be directly monitored using Lemon since it is not possible to install any monitoring agent in them. On the other hand, OEM (Oracle Enterprise Manager) provides integrated OEM Connectors that use SNMP to monitor NAS filers from an OEM agent running in a server node. OEM monitors performance as well as raises alerts and provides history log information.

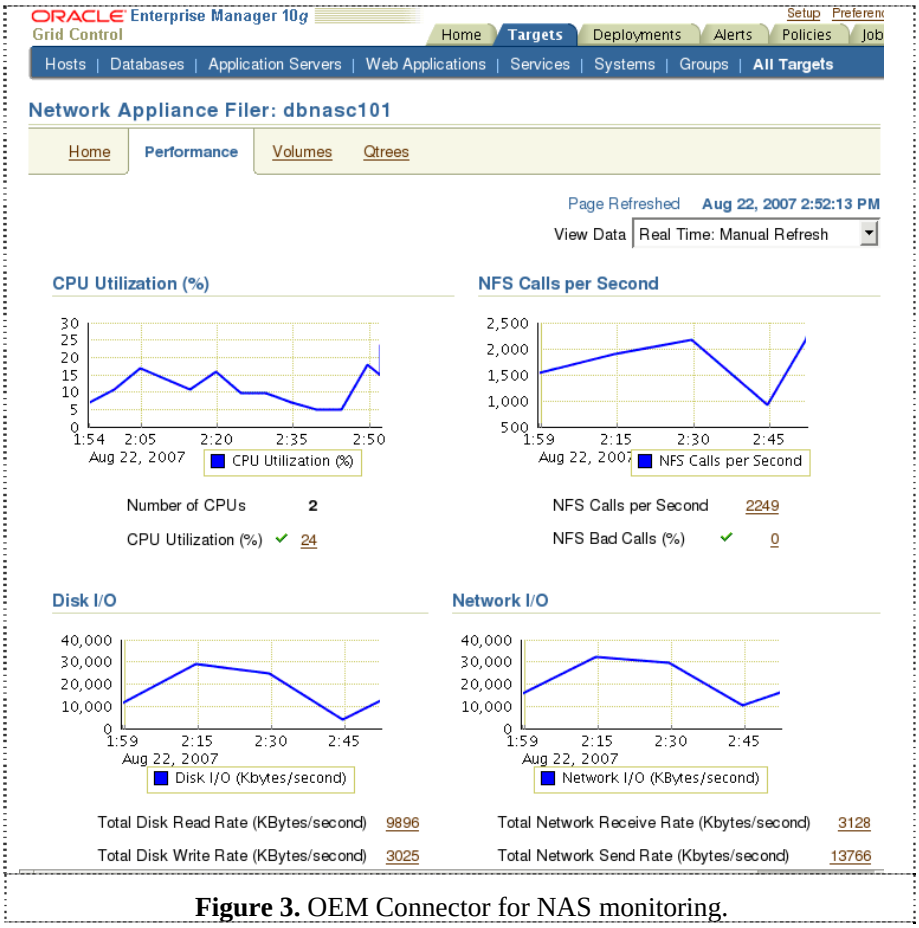

Figure 3. OEM Connector for NAS monitoring.

Additionally, Data ONTAP offers several Auto Support options to allow automatic report of hardware problems to the Hardware Support Centre. This is done via SMTP (it can as well be done via HTTPS). Since NAS filers are not connected to the general-purpose network, SMTP communication goes through one of the Server nodes.

7. Conclusion

Over 15 Oracle RAC instances are already working in production using this NAS based infrastructure. This includes RAC services for various CASTOR 2 Stagers (for CMS, Atlas, Alice, LHCb, etc) as well as other projects at CERN (Lemon, OEM) and some AIS services.

A set of hardware failure tests has been performed. Tests included:

- Powering off each network switch
- Powering off NAS filer
- Powering off Oracle RAC member node

- Unplugging interconnect active bond member network cable
- Unplugging NAS filer active member VIF cable
- Unplugging FC cable connecting NAS filer and Disk shelf
- Unplugging power cable of a NAS filer
- Unplugging power cable of Disk Shelf
- Removing a Disk from shelf

All these tests passed successfully and did not cause any service interruption. Also, several maintenance operations were performed causing no downtime:

- Failed Disks replacement
- NAS filer HA (Host Adapter) replacement
- Filer OS update Data ONTAP upgrade to 7.2.2 (rolling forward)
- Failed Power Supply Unit replacement (both in Server node and NAS Filer)
- Ethernet Network cable replacement
- Server Node OS reinstallation
- Server Nodes Kernel upgrade

As result of this very positive experience, the NAS based infrastructure will be extended with the objective to base all Oracle RAC installations on the same technology. In this way, IT/DES hopes to drastically reduce the burden of having to administrate and maintain the current rather fragmented setup based on a variety of suppliers and Store technologies (EMC, Sun Storage, etc). Moreover, this solution avoids the complexity presents in other solutions like SAN systems, which requires installing Oracle ASM or using a volume manager.

Acknowledgements

The authors would like to thank all IT-DES group for their involvement in the design and set up of this NAS based Database infrastructure. We would like to particularly thank Eric Grancher, Nilo Segura Chinchilla, Artur Wiecek, Mats Möller, Johan Gudheim Hansen and Philippe Defert for providing ideas and feedback on this project.

References

[1] Bonding on Linux: http://linux-net.osdl.org/index.php/Bonding
[2] ORACLE 10g Release 2: http://www.oracle.com/pls/db102/homepage
[3] ONTAP documentation 7.1.1.1